Fully Convolutional Crowd Counting on Highly Congested Scenes

Mark Marsden, Kevin McGuinness, Suzanne Little and Noel E. O'Connor
Insight Centre for Data Analytics, Dublin City University, Dublin, Ireland

Keywords: Computer Vision, Crowd Counting, Deep Learning.

Abstract: In this paper we advance the state-of-the-art for crowd counting in high density scenes by further exploring the idea of a fully convolutional crowd counting model introduced by (Zhang et al., 2016). Producing an accurate and robust crowd count estimator using computer vision techniques has attracted significant research interest in recent years. Applications for crowd counting systems exist in many diverse areas including city planning, retail, and of course general public safety. Developing a highly generalised counting model that can be deployed in any surveillance scenario with any camera perspective is the key objective for research in this area. Techniques developed in the past have generally performed poorly in highly congested scenes with several thousands of people in frame (Rodriguez et al., 2011). Our approach, influenced by the work of (Zhang et al., 2016), consists of the following contributions: (1) A training set augmentation scheme that minimises redundancy among training samples to improve model generalisation and overall counting performance; (2) a deep, single column, fully convolutional network (FCN) architecture; (3) a multi-scale averaging step during inference. The developed technique can analyse images of any resolution or aspect ratio and achieves state-of-the-art counting performance on the Shanghaitech Part_B and UCF_CC_50 datasets as well as competitive performance on Shanghaitech Part_A.

1 INTRODUCTION

Vision based crowd size estimation, often referred to as crowd counting, has become an important topic for the computer vision community. Crowd counting algorithms attempt to produce an accurate estimation of the true number of people present in a crowded scene. A crowd count is inherently more objective than other crowd size representations (e.g. crowd density level) but is also more challenging to produce. Accurate knowledge of the crowd size in a public space can provide valuable insight for tasks such as city planning, analysing consumer shopping patterns as well as maintaining general crowd safety. Several key challenges such as visual occlusions and high levels of variation in scene content have limited progress in this area. Techniques developed for crowd counting can also be applied to tasks from other domains such as counting bacteria or cells in microscopic images (Xie et al., 2016).

Related work. Existing approaches to crowd counting largely fall into two categories: counting by detection and counting by regression.

Counting by detection approaches involve training a visual object detector to find and count each person in the scene. Each human is assumed to be an individual entity that must be found. These algorithms (Wu and Nevatia, 2005; Lin et al., 2001; Ge and Collins, 2009) are computationally demanding, requiring the image to be exhaustively analysed at multiple scales due to perspective issues, which alter the size of people in different parts of the scene. The robustness of these object detectors also suffers significantly due to visual occlusions, resulting in rapid performance degradation as a crowd becomes highly congested (i.e. several hundred people in frame).

Counting by regression techniques (Change Loy et al., 2013; Chen et al., 2012; Lempitsky and Zisserman, 2010; Liu and Tao, 2014; Chan and Vasconcelos, 2012) on the other hand attempt to learn a direct mapping between low-level features and the overall number of people in frame or within a frame region. Individual people are not explicitly detected or tracked in these approaches, meaning visual occlusions have less impact on counting accuracy. While generally more computationally efficient than counting by detection methods, regression-based techniques have suffered greatly from overfitting in the past due to a lack of varied training data. To remedy this, a number of high density, high variation crowd counting datasets such

Marsden M., McGuinness K., Little S. and E. OâĂŹConnor N.
Fully Convolutional Crowd Counting on Highly Congested Scenes.
DOI: 10.5220/0006097300270033
In *Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2017)*, pages 27-33
ISBN: 978-989-758-226-4

as UCF_CC_50 (Idrees et al., 2013) and Shanghaitech (Zhang et al., 2016) have emerged. Recent advancements in graphical processing unit (GPU) hardware and the availability of very large, labelled datasets such as ImageNet (Deng et al., 2009) have resulted in deep learning approaches such as convolutional neural networks (CNN) achieving state-of-the-art performance in many computer vision tasks (image classification, face detection, object detection). Deep learning techniques have recently been applied to the task of regression-based crowd counting (Zhang et al., 2015; Hu et al., 2016; Zhang et al., 2016), resulting in a notable improvement in counting accuracy, especially for high density scenes (i.e. where there are 1000+ people in frame).

Fully convolutional networks (FCN) are a unique variation on the CNN technique where a proportionally sized feature map output is produced for a given input image rather than a classification label or regression score. FCNs have been used for a variety of tasks including semantic segmentation (Long et al., 2015) and saliency prediction (Pan et al., 2016). Zhang et al. (Zhang et al., 2016) trained an FCN to transform an image of a crowded scene into a crowd density heatmap, which when integrated produces a highly accurate crowd count estimate, even for very challenging scenes. One of the key aspects of fully convolutional nets that makes the method particularly suited to crowd counting is the use of a variable size input, allowing the model to avoid the loss of detail and visual distortions typically encountered during image downsampling and reshaping.

Contributions of this Paper. The core objective of this paper is to achieve highly accurate crowd counting on densely congested scenes. This study will further explore the idea of a fully convolutional crowd counting model originally introduced by (Zhang et al., 2016). The core contributions can be summarized as follows:

1. A training set augmentation scheme is proposed which minimises redundancy among training samples in order to improve model generalisation and overall counting performance.
2. A deep, single column, fully connected network is used to generate crowd density heatmaps. The greater model capacity improves the FCNs ability to learn the highly abstract, nonlinear relationships present in crowd counting datasets.
3. To overcome the scale and perspective issues that often limit the accuracy of crowd counting algorithms a given test image is fed into the network at multiple scales (e.g. original size + 80% original size). The crowd count is estimated for each scale and the mean is taken as the overall estimate.

Table 1: Shanghaitech Part_B validation performance using different training set augmentation schemes. Horizontal flips are used in all cases.

Augmentation Scheme	**MAE**	**MSE**
None	30.5	47.5
4 quadrants crops + 1 overlapping centre crop	25.5	36.5
4 quadrants crops	**24.1**	**33.5**

This simple step taken during inference results in significant performance gains.

2 A FULLY CONVOLUTIONAL NETWORK FOR CROWD COUNTING

A fully convolutional network (FCN) allows for the input images used during training and inference to be of any resolution and aspect ratio, thanks to the absense of any fully connected layers. Rather than produce a fixed size classification or regression output, FCNs generate a feature map or set of feature maps proportionally sized to the input image. This type of network can then be used for a range of tasks including image transformation and pixel wise regression/classification (Pan et al., 2016; Long et al., 2015).

Zhang et al. (Zhang et al., 2016) trained an FCN to transform an image of a crowded scene into a crowd density heatmap, which when integrated produces a highly accurate crowd count estimate. In order to train a network to produce this function a set of ground truth heatmap images must be generated for which the integral is equal to the pedestrian count. The head annotations found in most crowd counting datasets can be used to this end. For each of the N head annotations associated with a given training image a unit impulse is added to the heatmap ground truth at the given location, as described in equation 1 where x_i is the position of a given head.

$$H(x) = \sum_{i=0}^{N} \delta(x - x_i) \tag{1}$$

.

To convert this discrete density heatmap to a continuous function, convolution with an adaptive Gaussian kernel $G_{\sigma i}$ is applied for each head annotation (Zhang et al., 2016). The spread parameter σ used for a given head annotation x_i is decided based on the mean distance to the 5 nearest heads $\bar{d}_i$, using equation

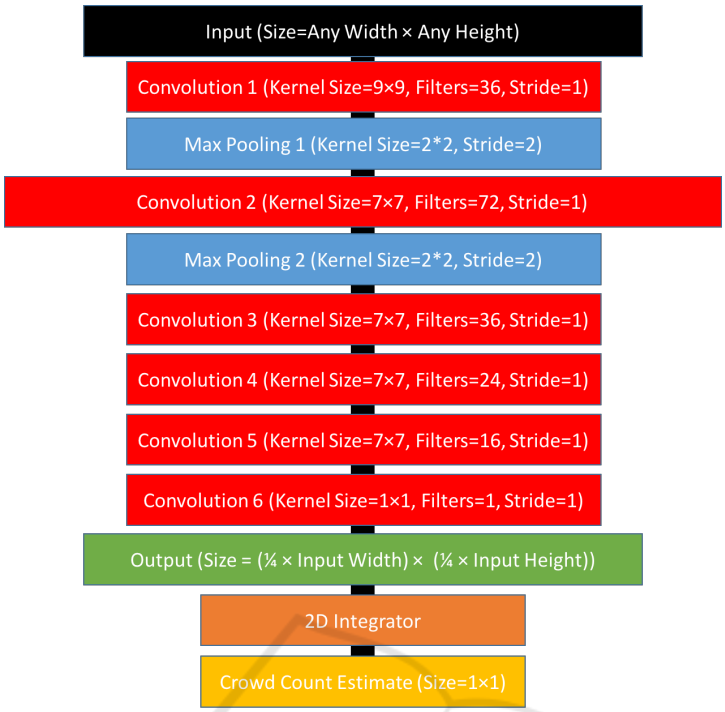

Figure 1: Fully convolutional network architecture used to perform crowd counting. Each convolutional layer is followed by a ReLU activation layer apart from "Convolution 6". A 2D integration (simply an element-wise sum in this case) is applied to the network output in order to produce the crowd count estimate value.

2. Distance to the surrounding heads roughly correlates with proximity to the camera, producing more smoothing the closer to the camera a pedestrian is, helping us account for perspective distortion issues. The 0.3 weighting was found empirically by (Zhang et al., 2016) to produce optimal results and is maintained. This fully convolutional approach to crowd counting will form the basis of our technique.

$$G_{\sigma i} = 0.3 * \bar{d}_i \quad (2)$$

2.1 Training Set Augmentation Scheme

The training set generation scheme used and particularly the chosen augmentation techniques, play an important role in the strong counting accuracy achieved by our method. Most crowd counting datasets consist of only a few hundred images, making augmentation an essential step. Taking several image crops to increase training set size and variation is a common augmentation technique used in computer vision. While it is perfectly acceptable to allow these crops to overlap for image recognition tasks, pixel-wise tasks can potentially overfit when the network is continually exposed to a given set of pixels during training. Therefore our augmentation scheme is developed to ensure there is no such redundancy. For each training set image the four image quadrants as well as their horizontal flips are taken as training samples, ensuring no overlap. In order to validate this augmentation scheme the Shanghaitech Part_B training set is further split into training and validation subsets using a 9:1 ratio. Table 1 highlights the difference in validation performance when our model is trained on a dataset with and without overlapping crops. Both runs are trained from scratch using the same network architecture. This simple change results in a notable improvement in counting accuracy, despite the reduction in overall training set size.

2.2 FCN Architecture

Processing high resolution images (e.g. 1000×1000 pixels) using a fully connected network presents certain challenges and constraints, particularly in terms of memory usage on GPU hardware. We are limited in the number of convolutional kernels and layers (i.e. model capacity) our FCN can have. Therefore we must attempt to design the best possible FCN architecture capable of processing high resolution images such as those in the UCF_CC_50 dataset. An Nvidia GTX 970 card with 4GB of VRAM was used for our experiments. With these constraints in mind we designed a 6 layer, single column FCN as illustrated in figure 1. This network contains just 315,000 parameters, thanks

Table 2: Shanghaitech Part_B validation performance using different network architectures.

Network Architecture	MAE	MSE
Proposed	**24.1**	**33.5**
Multi-Column FCN (Zhang et al., 2016)	25.5	36.5

largely to the absence of any fully connected layers. Rectified linear unit (ReLU) activations are applied after each convolutional layer apart from the last. 1×1 convolutions are used in the final layer to produce the single channel crowd density heatmap. This density heatmap is then fed into a 2D integrator (simply an element-wise sum in this case) to produce the crowd count estimate. The network is optimised in a single training run using stochastic gradient descent and backpropagation. We chose to minimise the Euclidean distance between the produced density heatmap and the ground truth heatmap. This loss function is fully defined as follows:

$$L(\Theta) = \frac{1}{2N} \sum_{i=1}^{N} \|F(X_i;\Theta) - F_i\|_2^2, \qquad (3)$$

where Θ is the set of network parameters to optimise, N is the batch size, X_i is the i^{th} batch image and $\mathrm{F_i}$ is the corresponding ground truth density heatmap. $F(X_i;\Theta)$ is the estimated heatmap for a given batch image X_i.

Table 2 presents the difference in Shanghaitech Part_B validation performance when our high capacity architecture is used over a shallower multi-column FCN architecture (Zhang et al., 2016). All hyperparamaters including the training set augmentation scheme are kept identical for both runs. We can see a clear improvement in performance when our deeper single column architecture is used.

2.3 Multi-Scale Averaging During Inference

Scale and perspective issues often limit the performance of crowd counting algorithms. A top down camera perspective is ideal for this task but cannot be guaranteed in real world settings. In most CCTV scenarios foreground pedestrians are much larger than those in the background, who may only occupy a few pixels. As FCNs allow for a variable size input image, we can easily resize a given test image before feeding it into the network and estimating the crowd size. A scaled down version may result in more accurate crowd counting in certain scene regions than the original. Therefore in order to overcome these issues a given test image is fed into the network at multiple scales (e.g. original size + 80% original size). The crowd count is estimated for each scale and the mean is taken as the overall estimate. Table 3 shows the validation performance of several multi-scale averaging schemes. The same training and validation subsets are used as before. Scheme 2 performs best and is thus used for all further experiments.

3 EXPERIMENTS

The performance of our fully convolutional crowd counting technique is evaluated on three crowd counting benchmarks from two datasets. These benchmarks vary greatly in terms of congestion level and scene content. Our model achieves very strong crowd counting performance, particularly on images of high density scenes with several thousand people in frame. The Caffe framework (Jia et al., 2014) and it's Python wrapper are used to train and deploy our model. Both mean absolute error (MAE) and mean squared error (MAE) are used to compare crowd counting performance on all datasets. These two metrics are defined as follows:

$$\mathrm{MAE} = \frac{1}{N} \sum_{i=1}^{N} |z_i - \check{z}_i|, \qquad (4)$$

$$\mathrm{MSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (z_i - \check{z}_i)^2}, \qquad (5)$$

where N is the number of test images, z_i is the actual number of people in the i^{th} image and $\check{z}_i$ is the estimated number of people in the i^{th} image. MAE indicates the accuracy of the estimates while MSE corresponds to the robustness of the estimates.

3.1 Shanghaitech Dataset

The Shanghaitech dataset (Zhang et al., 2016) contains 1198 images of crowded scenes with a total of 330,165 head annotations included. This dataset is split into two parts; Part_A contains images of high density scenes (up to 3000 people) taken from the internet while Part_B consists of medium density crowd images (up to 600 people) taken in the streets of Shanghai. Each part consists of a respective training and test set. The performance of the proposed approach is evaluated on each part separately.

The redundancy minimising augmentation approach discussed in section 2.1 is used for both parts. The network is trained from scratch in a single run for $2e^6$ iterations using a base learning rate of $1e^{-6}$, with the learning rate decreased by a factor of 10 after $1e^6$

Table 3: Shanghaitech Part_B validation performance using different mutli-scale averaging inference schemes.

Multi-scale Averaging Scheme	MAE	MSE
1) None	24.1	33.5
2) MeanCount(Original Size, 80% Original Size)	**22.1**	**31.5**
3) MeanCount(Original, 80% Original Size, 70% Original Size)	24.6	34.1
4) MeanCount(Original, 80% Original Size, 70% Original Size, 60% Original Size)	25.2	34.8

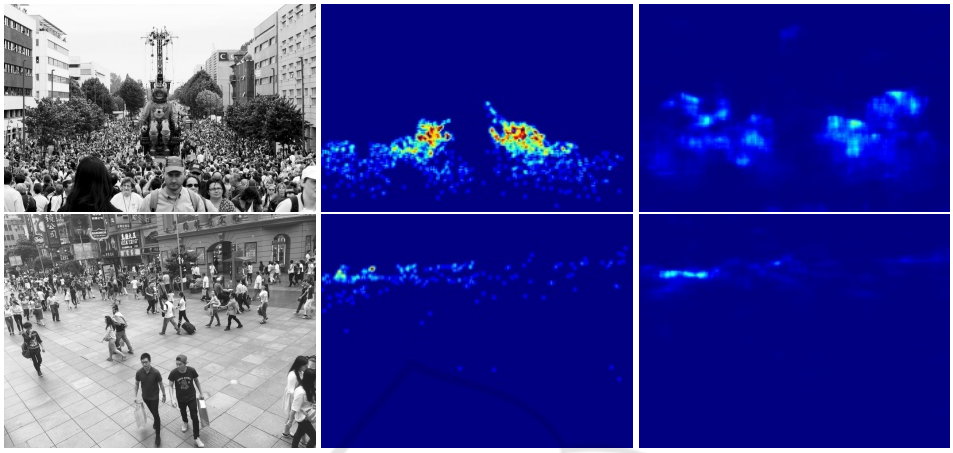

Figure 2: **Top:** Shanghaitech Part_A test image, Ground Truth Heatmap, Estimated Heatmap. True Count=1326, Estimated Count=1138. **Bottom:** Shanghaitech Part_B test image, Ground Truth Heatmap, Estimated Heatmap. True Count=240, Estimated Count=242.

iterations. Gaussian weight initialisation with a standard deviation of 0.01 is used as well as a weight decay of 0.0005 and a momentum of 0.9. Due to memory limitations and the high image resolution of the Shanghaitech dataset a batch size of 1 is used during training. During testing the proposed multi-scale averaging is applied, using scheme 2 from table 3.

Our method is compared to the existing approaches in table 4 and achieves state-of-the-art performance on Shanghaitech Part_B, improving MAE by 10% and MSE by 19%. Competitive performance is also achieved on Shanghaitech Part_A, with an MSE near identical to the state-of-the-art produced. Figure 2 shows our technique in action on images from this dataset.

3.2 UCF_CC_50 Dataset

The UCF_CC_50 dataset (Idrees et al., 2013) contains 50 highly challenging crowd images taken from the Internet. The number of pedestrians present in a frame ranges between 94 and 4500. Following convention (Idrees et al., 2013) a 5-fold cross validation is performed on this dataset. The same augmentation process, training hyperparamaters and multi-scale averaging scheme are used as for the Shanghaitech dataset. Again due to memory limitations a batch size of 1 is used. Table 5 compares our technique with the existing approaches, with our method improving the state-of-the-art for MAE and MSE by 11% and 13% respectively. Figure 3 shows our technique in action on an image from this dataset.

3.3 Cross Dataset Performance

In order to investigate the generalisation potential of our technique, we performed a number of cross dataset experiments. In each experiment we take a model trained on a specific dataset as the ”source” domain and then evaluate MAE and MSE performance on another unseen dataset or ”target” domain. The results of these experiments are shown in Table 6. Superior performance is achieved when our source and target domain both contain images of a similar density level (Shanghaitech Part_B => Shanghaitech Part_A, Shanghaitech Part_A => UCF_CC_50). On the other hand very poor performance is achieved when our source domain contains significantly higher density images than the target (UCF_CC_50 => Shanghaitech Part_A). Therefore a model used for real world deployment must be trained on an appropriately large and varied training set.

Table 4: Comparing the performance of different crowd counting approaches on the Shanghaitech dataset.

	Part_A		Part_B	
Method	**MAE**	**MSE**	**MAE**	**MSE**
(Zhang et al., 2016)	**110.2**	**173.2**	26.4	41.3
(Zhang et al., 2015)	181.8	277.7	32.0	49.8
Proposed Approach	126.5	173.5	**23.76**	**33.12**

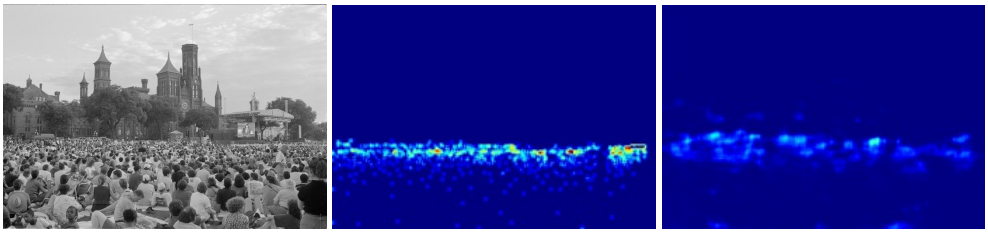

Figure 3: UCF_CC_50 test image, Ground Truth Heatmap, Estimated Heatmap. True Count=1544, Estimated Count=1566.

Table 5: Comparing performance of different crowd counting approaches on the UCF_CC_50 dataset.

Method	**MAE**	**MSE**
(Rodriguez et al., 2011)	655.7	697.8
(Lempitsky and Zisserman, 2010)	493.4	487.1
(Idrees et al., 2013)	419.5	541.6
(Zhang et al., 2016)	377.6	509.1
(Hu et al., 2016)	431.5	438.5
(Zhang et al., 2015)	467.0	498.6
Our Approach	**338.6**	**424.5**

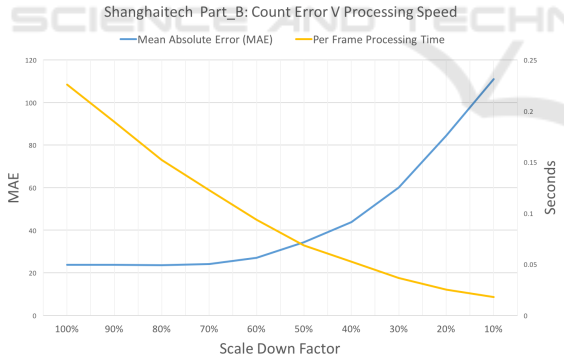

Figure 4: Comparing count error and processing speed as we scale down the image resolution on Shanghaitech Part_B.

3.4 Trade-off between Computation Speed and Counting Accuracy

The ability of a fully convolutional network to process images of any resolution is one of the key reasons behind the strong counting performance achieved by our method. However, analysing such high resolution images results in high memory consumption and slower processing speed during inference. Therefore we want to investigate to what degree image resolution can be reduced during test time before we see significant performance degradation. Doing so we can find the best possible trade-off between computation speed and accuracy. The Shanghaitech Part_B dataset is used for this experiment. Test images are scaled down to a given percentage of their original size with aspect ratio maintained. The results of these experiments are presented in figure 4. Surprisingly we do not see the error increase significantly until we reduce the image size to 50%. However with this 50% resizing applied the processing speed is increased by a factor of 4. In deployment scenarios this type of downsampling can be applied in order to analyse real-time video without a major loss of accuracy.

4 CONCLUSION

In this paper we have proposed a deep, fully convolutional crowd counting model that can perform highly accurate single image crowd counting in almost any surveillance scenario. Our model achieves state-of-the-art performance on both the Shaghaitech Part_B and UCF_CC_50 datasets as well as competitive performance on the Shaghaitech Part_A dataset. Images of any resolution and aspect ratio can be analysed. The developed approach also performs well even with significant image downsampling applied at test time. Future work in this area will look to extend our network to also perform other pixel-wise tasks such as crowd segmentation in order to exploit the inter-task correlations present.

Table 6: Cross dataset performance of our method. The percentage increases in MAE and MSE are highlighted.

Source Domain	Target Domain	MAE	MSE
Shanghaitech_B	Shanghaitech_A	191(+52%)	337.5(+94%)
UCF_CC_50	Shanghaitech_A	269(+116%)	359.5(107%)
Shanghaitech_A	Shanghaitech_B	68(+189%)	100.5(+200%)
UCF_CC_50	Shanghaitech_B	165(+614%)	215(+540%)
Shanghaitech_A	UCF_CC_50	473(+40%)	680(+50%)
Shanghaitech_B	UCF_CC_50	699(+100%)	866 (+105%)

ACKNOWLEDGMENTS

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/12/RC/2289.

REFERENCES

Chan, A. B. and Vasconcelos, N. (2012). Counting people with low-level features and bayesian regression. *IEEE Transactions on Image Processing*, 21(4):2160–2177.

Change Loy, C., Gong, S., and Xiang, T. (2013). From semi-supervised to transfer counting of crowds. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2256–2263.

Chen, K., Loy, C. C., Gong, S., and Xiang, T. (2012). Feature mining for localised crowd counting. In *BMVC*, volume 1, page 3.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. (2009). ImageNet: A Large-Scale Hierarchical Image Database. In *CVPR09*.

Ge, W. and Collins, R. T. (2009). Marked point processes for crowd counting. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2913–2920. IEEE.

Hu, Y., Chang, H., Nian, F., Wang, Y., and Li, T. (2016). Dense crowd counting from still images with convolutional neural networks. *Journal of Visual Communication and Image Representation*, 38:530–539.

Idrees, H., Saleemi, I., Seibert, C., and Shah, M. (2013). Multi-source multi-scale counting in extremely dense crowd images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2547–2554.

Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., and Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM international conference on Multimedia*, pages 675–678. ACM.

Lempitsky, V. and Zisserman, A. (2010). Learning to count objects in images. In *Advances in Neural Information Processing Systems*, pages 1324–1332.

Lin, S.-F., Chen, J.-Y., and Chao, H.-X. (2001). Estimation of number of people in crowded scenes using perspective transformation. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 31(6):645–654.

Liu, T. and Tao, D. (2014). On the robustness and generalization of cauchy regression. In *2014 4th IEEE International Conference on Information Science and Technology*, pages 100–105. IEEE.

Long, J., Shelhamer, E., and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3431–3440.

Pan, J., McGuinness, K., Sayrol, E., O'Connor, N., and Giroi Nieto, X. (2016). Shallow and deep convolutional networks for saliency prediction.

Rodriguez, M., Laptev, I., Sivic, J., and Audibert, J.-Y. (2011). Density-aware person detection and tracking in crowds. In *2011 International Conference on Computer Vision*, pages 2423–2430. IEEE.

Wu, B. and Nevatia, R. (2005). Detection of multiple, partially occluded humans in a single image by bayesian combination of edgelet part detectors. In *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, pages 90–97. IEEE.

Xie, W., Noble, J. A., and Zisserman, A. (2016). Microscopy cell counting and detection with fully convolutional regression networks. *Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization*, pages 1–10.

Zhang, C., Li, H., Wang, X., and Yang, X. (2015). Cross-scene crowd counting via deep convolutional neural networks. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 07-12-June, pages 833–841.

Zhang, Y., Zhou, D., Chen, S., Gao, S., and Ma, Y. (2016). Single-image crowd counting via multi-column convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 589–597.